Probability Theory

Zambelli Lorenzo
BSc Applied Mathematics

April 2021-June 2021

1 Introduction

This summaries has been made primly by used the lecture notes [1] with the help of [2] and some websites.

1.1 Element of Set Theory

1. **Cardinality:** numbers of elements of a finite set (its often denoted by $\#(A)$).
2. Given subsets $A_1, A_2, ...$ of M their **union** $\bigcup_{j=1}^{\infty} A_j$ and their **intersection** $\bigcap_{j=1}^{\infty} A_j$ is the set of those $x \in M$ that belong to a least one of the A_j or that belong to all A_j, respectively
3. **Distributive Law:**
$$A \cap \left(\bigcup_{j=1}^{\infty} B_j \right) = \bigcup_{j=1}^{\infty} (A \cap B_j)$$
4. Two sets A and B are said to be **disjoint** provided that $A \cap B = \emptyset$.
5. A sequence of sets $A_1, A_2, ...$ is called **pairwise disjoint** whenever $A_i \cap A_j = \emptyset$ if $i \neq j$
6. The **complementary set** of $B \subseteq M$ is $B^c := \{x \in M : x \notin B\}$
7. Let $A, B \subseteq M$, then the **difference** $A \backslash B$ is defined by $\{w \in M : w \in A \text{ and } w \notin B\}$ or, similarly $A \backslash B = A \cap B^c$
8. **De Morgan's rules**
$$\left(\bigcup_{j=1}^{\infty} A_j \right)^c = \bigcap_{j=1}^{\infty} A_j^c \qquad \left(\bigcap_{j=1}^{\infty} A_j \right)^c = \bigcup_{j=1}^{\infty} A_j^c$$

1.2 Combinatorics

1.2.1 The rules of sum and product

The Rule of Sum and Rule of Product are used to decompose difficult counting problems into simple problems.

- **Rule of Sum:** If a sequence of tasks $T_1, T_2, ..., T_m$ can be done in $w_1, ..., w_m$ ways respectively (no tasks can be performed simultaneously), then the number of ways to do one of these task is $\sum_{j \geq 1} w_j$, i.e if we consider two task A and B which are disjoint, then $\#(A \cap B) = \#(A) + \#(B)$

- **Rule of Product:** If we have set of events $A_1, A_2, ...$ where A_1 occur before A_2, A_2 occur before A_3, and so on, then $\#\left(\prod_{j\geq 1} A_j\right) = \prod_{j\geq 1} \#(A_j)$

1.2.2 Permutations

A **permutation** is an arrangement of some elements in which order matters. In other words a Permutation is an ordered Combination of elements.

1. Let S_n be the set of all permutations of order n. Then how many ways I can rearrange the numbers $\{1, ..., n\}$?
$$\#(S_n) = n!$$
or, equivalently, there are $n!$ different ways to order n distinguishable objects

2. How many ways are there to pick a sequence of k (not necessarily distinct) numbers chosen from $1, ..., n$?
$$n^k$$

3. How many ways are there to pick a sequence of k distinct numbers chosen from $1, ..., n$? in this case makes sense if $k \leq n$
$$(n)_k = \frac{n!}{(n-k)!}$$

4. How many ways can you distribute n objects into one group of k and into another of $n-k$ elements? or How many subsets of $\{1, ..., n\}$ of cardinality exactly k are there?
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$
This is called **the binomial coefficients**, read "n chosen k"

5. How many ways are there to draw k balls out of $1, ..., n$ with replacement but without order?
$$\binom{n+k-1}{k}$$

6. How many ways you can distribute n elements into m groups of sizes $k_1, k_2, ..., k_m$ where $k_1 + \cdots + k_m = n$?
$$\binom{n}{k_1, ..., k_m} := \frac{n!}{k_1! \cdots k_m!}, \quad k_1 + \cdots + k_m = n$$
This is called **multinomial coefficient**, read "n chosen k_1 up to k_m"

7. **Pascal's triangle**
$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

1.2.3 Inclusion-Exclusion principle

The Inclusion-exclusion principle computes the cardinal number of the union of multiple non-disjoint sets. For two sets A and B, the principle states

$$\#(A \cup B) = \#(A) + \#(B) - \#(A \cap B)$$

the generalized formula

$$\#\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{1\leq i<j<k\leq n} \#(A_i \cap A_j) + \sum_{1\leq i<j<k\leq n} \#(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1} \#(A_1 \cap \cdots \cap A_2)$$

2 Probability Spaces

The basic concern of Probability Theory is to model experiments involving randomness, that is, experiments with nondetermined outcomes, shortly called random experiments.

Definition 2.1 *Random experiments are described by probability spaces* $(\Omega, \mathbb{A}, \mathbb{P})$

Definition 2.2 ***The sample space*** Ω *is a nonempty set that contains (at least) all possible outcomes of the random experiment.*

Remark: Due to mathematical reasons sometimes it can be useful to choose K larger than necessary. It is only important that the sample space contains all possible results.

Definition 2.3 *Any element* $w \in \Omega$ *is called an* ***outcome****. Any subset* $A \subseteq \Omega$ *is called* ***event***

Definition 2.4 *A* ***Probability space*** *is a triple* $(\Omega, \mathcal{A}, \mathbb{P})$*, where* Ω *is a sample space,* $\mathcal{A}$ *is*

$$\mathcal{A} = \begin{cases} \text{set of all subsets of } \Omega & \text{if } \Omega \text{ is countable} \\ \text{a certain set of subsets of } \Omega & \text{if } \Omega \text{ is uncountable} \end{cases}$$

and $\mathbb{P} : \mathcal{A} \to [0, 1]$ *is a probability measure (probability function)*

Definition 2.5 *Let* Ω *be a sample space and let* $\mathcal{A}$ *be as in definition 2.4. A function* $\mathbb{P} : \mathcal{A} \to [0, 1]$ *is called* ***probability measure*** *on* $(\Omega, \mathcal{A})$ *if*

1. $\mathbb{P}(\emptyset) = 0$ *and* $\mathbb{P}(\Omega) = 1$

2. *if* $A_1, A_2, ...$ *are pairwise disjoint, then*

$$\mathbb{P}\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} \mathbb{P}(A_i)$$

This are called ***Kolmogorov's axioms of probability****. (2) is often called sigma-additivity.*

Theorem 2.6 ***First properties of probabilities***

- $\mathbb{P}(\emptyset) = 0$
- *if* $A, B \in \Omega$ *satisfy* $A \subseteq B$*, then* $\mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ *for any* $A \in \Omega$
- $\mathbb{P}(A) \leq 1$ *for any* A
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *(Inclusion-Exclusion principle)*
- *Probability measures are* ***monotone****, that is, if* $A \subseteq B$ *then* $\mathbb{P}(A) \leq \mathbb{P}(B)$

Theorem 2.7 ***Sigma sub-additivity*** *For any* $A_1, A_2, ...,$ *not necessarily disjoint*

$$\mathbb{P}\left(\bigcup_{i\geq 1} A_i\right) \leq \sum_{i\geq 1} \mathbb{P}(A_i)$$

Theorem 2.8 ***Inclusion-exclusion formula*** *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space and let* $A_1, A_2, ..., A_n$ *be some (not necessarily disjoint events, then*

$$\mathbb{P}\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\leq j_1 < \cdots < j_k \leq n} \mathbb{P}(A_{j_1} \cap \cdots \cap A_{j_k})$$

Theorem 2.9 *Suppose* $\Omega = \{w_1, w_2, ...\}$ *and* $p_1, p_2, ...$ *are non negative numbers with* $\sum p_i = 1$*. Defining, for all* $A \in \Omega$*,*

$$\mathbb{P}(A) = \sum_{i:w_i \in A} p_i$$

Then $\mathbb{P}$ *is a probability measure.*

Theorem 2.10 *If* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space,* Ω *is finite and the outcomes* $w \in \Omega$ *all have the same probability, then, for any* $A \in \mathcal{A}$*,*

$$\mathbb{P}(A) = \frac{\#(A)}{\#(\Omega)}$$

3 Conditional probability and independence

Definition 3.1 *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space and* A, B *events; assume* $\mathbb{P}(B) > 0$*. Then the probability of* A *given* B *is defined by*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Definition 3.2 *The mapping* $\mathbb{P}(\dot{|}B)$ *is called* ***conditional probability*** *or also* ***conditional distribution*** *(under condition* B*)*

Remark: The main advadge of this definition is that it implies that conditional probabilities share all the proprieties of ordinary probability measures.

Theorem 3.3 ***Law of total probability*** *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space and let* $B_1, ..., B_n$ *in* $\mathcal{A}$ *be disjoint with* $\mathbb{P}(B_j) > 0$ *and* $\bigcup_{j=1}^n B_j = \Omega$. *Then for each* $A \in \mathcal{A}$ *holds*

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(B_j)\mathbb{P}(A|B_j)$$

Theorem 3.4 ***Bayes'formula*** *Suppose we are given disjoint events* B_1 *to* B_n *satisfying* $\bigcup_{j=1}^n B_j = \Omega$ *and* $\mathbb{P}(B_j) > 0$. *Let* A *be an event with* $\mathbb{P}(A) > 0$. *Then for each* $j \leq n$ *the following equation holds:*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\mathbb{P}(A|B_j)}{\sum_{i=1}^n \mathbb{P}(B_i)\mathbb{P}(A|B_i)}$$

Remark: in case $\mathbb{P}(A)$ is already known, Bayes's formula simplifies to

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\mathbb{P}(A|B_j)}{\mathbb{P}(A)}, \quad j = 1, ..., n$$

Definition 3.5 *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space. Two events* A *and* B *in* $\mathcal{A}$ *are said to be* ***independent*** *provided that*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

In the case that this eq. does not hold, the events A *and* B *are said* ***dependent***

Definition 3.6 *Events* $A_1, ..., A_n$ *are said to be* ***pairwise independent*** *if, whenever* $i \neq j$, *then*

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$$

In other words, for all $1 \leq i < j \leq 1$ *the events* A_i *and* A_j *are independent*

Definition 3.7 *The events* $A_1, ..., A_n$ *are said to be* ***mutually independent*** *provided that for each subset of* $I \subseteq \{1, ..., n\}$ *we have*

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$$

Remark: If $A_1, ..., A_n$ are mutually independent then they are also pairwise independent. However, in general the converse does not hold.

4 Random Variables

The are two ways to model a random experiment. The classical approach is to construct a probability space that describes this experiment. Another way is to choose a random variable X so that the probability of the occurrence of an event $B \in \mathbb{R}$ equals $\mathbb{P}(X \in B)$.

Definition 4.1 *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space. A mapping* $X : \Omega \to \mathbb{R}$ *is called a (real-valued)****random variable***

Remark: $w \in \Omega$ such that $X(w) = x$ is a event.

4.1 Probability Distribution of a Random Variable

Suppose we are given a random variable $X : \Omega \to \mathbb{R}$. We define now a mapping $\mathbb{P}_X$ from $\mathcal{A}$ to $[0, 1]$ as follows:

$$\mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B)) = \mathbb{P}(w \in \Omega : X(w) \in B) = \mathbb{P}(X \in B)$$

Definition 4.2 *Two random variables X_1 and X_2 are said to be **identically distributed** provided that $\mathbb{P}_{X_1} = \mathbb{P}_{X_2}$. Hereby, it is not necessary that X_1 and X_2 are defined on the same sample space. Only their distributions have to coincide.*

Definition 4.3 *Let X be a random variable, either discrete or continuous. Then its **cumulative distribution function** $F_X : \mathbb{R} \to [0, 1]$ is defined by*

$$F_X(x) = \mathbb{P}(X \leq x)$$

Theorem 4.4 *The distribution function F_X of the random variable X satisfies:*

1. *$F_X(-\infty) = 0$ and $F_X(\infty) = 1$*
2. *F_X is nondecreasing*
3. *F_X is continuous from the right*

Lemma 4.5 $\mathbb{P}(X = x) = F_X(x) - \lim_{y \nearrow x} F_X(y)$.

4.1.1 Discrete Random Variable

Definition 4.6 *A random variable X is **discrete** provided there exists an at most countably infinite set $D \subset \mathbb{R}$ such that $X : \Omega \to D$.*
In other words, a random variable is discrete if it attains at most countably infinite many different values.

Remark: If a random variable X is discrete with values in $D \subset \mathbb{R}$, then $\mathbb{P}_X(D) = \mathbb{P}(X \in D) = 1$.

Without losing generality we may always assume the following: if a random variable X has a discrete probability distribution, that is, $\mathbb{P}(X \in D) = 1$ for some finite or countably infinite set D, then X attains values in D.

Definition 4.7 *Let X be a discrete random variable, then the **probability mass function** of X is defined as follows*

$$f_X(x) = \mathbb{P}(X = x) \quad \text{for all } x$$

Remark: Note that if X is discrete, then

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{y \leq x} \mathbb{P}(X = y) = \sum_{y \leq x} f_X(y)$$

4.1.2 Continuous random variables

Definition 4.8 *A random variable X is said to be* ***continuous*** *provided that its distribution $\mathbb{P}_X$ is a continuous probability measure. That is, $\mathbb{P}_X$ possesses a* ***probability density function****, or pdf. Or a X is continuous if its cumulative distribution function $F_X(x)$ is continuous, i.e $\mathbb{P}(X = x) = 0$.*

Definition 4.9 *A pdf function of a continuous random variable X is a function $f_X : \mathbb{R} \to [0, \infty)$ that satisfies*

$$\mathbb{P}(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy \quad \textit{for all } x \in \mathbb{R}$$

Remark: for all real numbers $a < b$

$$\mathbb{P}(a \leq X \leq b) = F_X(x) = \int_{a}^{b} f_X(y)\, dy$$

Remark: If X is continuous the following are equal

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X < b)$$

and $\mathbb{P}(-\infty X < \infty) = 1$

4.2 Function of random variable

Let $g : \mathbb{R} \to \mathbb{R}$ and let Y, X be random variables. Let $Y = g(X)$, then

Definition 4.10 ***cdf***

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \in g^{-1}((-\infty, y])$$

Remark: if X discrete, then $F_Y(y) = \sum_{x \in g^{-1}} f_X(x)$. Instead if its continuous the integral.

Theorem 4.11

1. *If $Y = g(X)$ and g is strictly increasing, then $F_Y(y) = F_X(g^{-1}(y))$*
2. *If g is strictly decreasing and X is continuous, then $F_Y(y) = 1 - F_X(g^{-1}(y))$*

Definition 4.12 ***pmf***

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \mathbb{P}(X \in g^{-1}(y)) = \sum_{x \in g^{-1}(y)} f_X(x)$$

Definition 4.13 ***pdf*** *Assume X has pdf f_X and $Y = g(X)$ with g differentiable and strictly increasing or decreasing. Then*

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

5 Expected Value and Variance

5.1 Expected Value

Definition 5.1 ***Expectation*** *The expectation (or expected value or mean) of a random variable X is*

$$\mathbb{E}[x] = \begin{cases} \sum_x x \cdot f_X(x) & \textit{if } X \textit{ is discrete} \\ \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx & \textit{if } X \textit{ has pdf } f_X \end{cases}$$

provided that the sum or integral exists.

Remark: Since $x_i\mathbb{P}(X = x_i) \geq 0$ and $xf_X(x) \geq 0$ (for continuous) for non-negative X, for those random variables the sum and the integral is always well-defined, but may be infinite.

Theorem 5.2 *For $g : \mathbb{R} \to \mathbb{R}$ and a random variable X,*

$$\mathbb{E}[g(X)] = \begin{cases} \sum_x g(x) \cdot f_X(x) & \textit{if } X \textit{ is discrete} \\ \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\,dx & \textit{if } X \textit{ has pdf } f_X \end{cases}$$

provided that the sum or integral exists.

Theorem 5.3 *if X is a random variable, $a, b, c \in \mathbb{R}$ and $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}(g_1(X)), \mathbb{E}(g_2(X))$ exist, then,*

1. $\mathbb{E}(ag_1(X) + bg_2(X) + c) = a\mathbb{E}(g_1(X)) + b\mathbb{E}(g_2(X)) + c$
2. *If $g_1 \geq 0$, then $\mathbb{E}(g_1(X)) \geq 0$*
3. *If $g_1 \geq g_2$, then $\mathbb{E}(g_1(X)) \geq \mathbb{E}(g_2(X))$*
4. *If $a \leq g_1(X) \leq b$, then $a \leq \mathbb{E}(g_1(X)) \leq b$*

Theorem 5.4 $\mathbb{E}$ ***through cdf***

1. *If X is a discrete random variable that only assumes values on $\{0, 1, 2, ...\}$, then*

$$\mathbb{E}[X] = \sum_{n \geq 0} (1 - F_X(n))$$

2. *If X is a continuous and non-negative random variable, then*

$$\mathbb{E}[X] = \int_o^{\infty} 1 - F_X(x)\,dx$$

5.2 Variance

Definition 5.5 *For a random variable X and an integer n, we define the variance of X:*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2].$$

where $\mu = \mathbb{E}[X]$.
The positive square root of $\mathrm{Var}(X)$ is called the standard deviation of X.

Remark: $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2$
Interpretation: The expected value μ of a random variable is its main characteristic. It tells us around which value the observations of X have to be expected. But it does not tell us how far away from μ these observation will be on average. Are they concentrated around μ or are they widely dispersed? This behavior is described by the variance. It is defined as the average quadratic distance of X to its mean value. If $\mathrm{Var}(X)$ is small, then we will observe realizations of X quite near to its mean. Otherwise it is likely to observe values of X fa away from its expected value.

Theorem 5.6 *If X is a random variable and a, b are constants,*

$$\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X)$$

6 Discrete distributions

6.1 Discrete uniform distribution

Let a, b integers, $a < b$. A random variable X follows a discrete uniform distribution with parameters a and b (abbreviated: $X \sim \mathrm{Unif}(a, b)$) if

$$f_X(x) = \frac{1}{b - a + 1}, \quad x = a, a + 1, ..., b$$

(in words: X is equally likely to be equal to any of the integer between (and including) a and b).

1. Expectation: $\mathbb{E}(X) = \frac{a+b}{2}$
2. Variance $\mathrm{Var}(X) = \frac{(b-a+1)^2-1}{12}$

6.2 Bernoulli distribution

Let $\mathbb{P} \in [0, 1]$. X follows a Bernoulli distribution with parameter p, that is, $X \sim \mathrm{Ber}(p)$ if

$$f_X(1) = p; \quad f_X(0) = 1 - p$$

1. $\mathbb{E}(X) = 0 \cdot (1 - p) + 1 \cdot p = p$
2. $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mu^2 = p - p^2 = p(1 - p)$

Remark: A Bernoulli trial is an experiment which results in success with probability p and failure with $1 - p$. X is then 1 when there is success and 0 when there is failure.

6.3 Binomial distribution

The sample space is $\Omega = \{0, 1, ..., n\}$ for some $n \geq 1$ and p is a real number with $0 \leq p \leq 1$

Definition 6.1 *The probability measure* $\mathrm{Bin}(n, p)$ *defined by*

$$\mathbb{P}(X = x) = f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, ..., n$$

where f_X is the pmf of X, is called ***binomial distribution*** *with parameters n and p*

Remark: if $A \subseteq \{0, 1, ..., n\}$, then

$$\text{Bin}_{n,p}(A) = \sum_{k \in A} \binom{n}{k} p^k (1-p)^{n-k}$$

1. **Expectation:** $\mathbb{E}(X) = np$
2. **Variance:** $\text{Var}(X) = np(1-p)$

Remark: The binomial distribution describes the following experiment. We execute n times independently the same experiment where each time either success or failure may appear. The success probability is p. Then $\text{Bin}(n, p)$ of $X = x$ is the probability to observe exactly x times success or, equivalently, $n - x$ times failure.

Theorem 6.2 ***Binomial theorem***

$$(a+b)^n = \sum_{0 \leq k \leq n} \binom{n}{k} a^{n-k} b^k$$

6.4 Geometric distribution

Suppose we perform Bernoulli trials with probability p of success until the first success is obtain. Let X be the numbers of trials needed for observe success for the first time. Then, for $x \in \{1, 2, ...\}$,

$$f_X(x) = \mathbb{P}(X = x) = (1-p)^{x-1} p$$

we say that X follows a geometric distribution with parameter p, $X \sim \text{Geo}(p)$

1. **Expectation:** $\mathbb{E}(X) = \frac{1}{p}$
2. **Variance:** $\text{Var}(X) = (1-p)/p^2$

6.5 Poisson distribution

Let $\lambda > 0$. A random variable X follows the Poisson distribution with parameter λ, $X \sim \text{Poi}(\lambda)$ if

$$f_X(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k = 0, 1, ...$$

Remark: note that $\sum_{x \geq 0} f_X(k) = 1$

1. **Expectation**: $\mathbb{E}[X] = \sum_{k \geq 0} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda$
2. **Variance**: $\text{Var}(X) = \lambda$

Poisson approximation to Binomial: this is an approximation which can be summarized by: $\text{Bin}(n, p)$ is close to $\text{Poi}(\lambda)$ when n is large, p is small and np is close to λ.

Proposition 6.3 *Assume $(p_n)_{n \in \mathbb{N}}$ is a sequence such that*

$$p_n \in [0, 1] \quad \textit{for each } n \textit{ and } \lim_{n \to \infty} np_n = \lambda > 0.$$

Then, for each $k \in \mathbb{N}$,

$$\underbrace{\binom{n}{k} p_n^k (1-p_n)^{n-k}}_{f_X(k) \text{ for } X \sim \mathrm{Bin}(n,p_n)} \xrightarrow{n\to\infty} \underbrace{\frac{\lambda^k}{k!} e^{-\lambda}}_{f_X \text{ for } X \sim \mathrm{Poi}(\lambda)}$$

Remark: The Poisson distribution describes experiments where the number of trials is big, but the single success probability is small.

7 Continuous distributions

7.1 Uniform distribution

$a, b \in \mathbb{R}$, $a < b$. A random variable X follows a continuous uniform distribution between a and b if it has pdf

$$f_X(x) = \frac{1}{b-a} \quad \text{if } x \in (a,b)$$

and it is called **uniform random variable**.

1. **Expectation:** $\mathbb{E}[X] = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{b+a}{2}$
2. **Variance:** $\mathrm{Var}(X) = \int_a^b \left(x - \frac{b+a}{2}\right)^2 \cdot \frac{1}{b-a}\, dx = \frac{(b-a)^2}{12}$

7.2 Exponential distribution

Idea: ”waiting time until next ....”

Remark: The exponential distribution plays an important role for the description of lifetimes. For instance, it is used to determine the probability that the lifetime of a component part or the duration of a phone call exceeds a certain time T ¿ 0. Furthermore, it is applied to describe the time between the arrivals of customers at a counter or in a shop.

Let $\lambda > 0$. A random variable X with pdf

$$X \sim \mathrm{Exp}(\lambda) \Rightarrow f_X(x) = \lambda e^{-x\lambda}, \quad x > 0$$

Remark: $F_x = \int_{-\infty}^x f_X(x)\, dx = 1 - e^{-x\lambda}$ if $x > 0$

1. **Expectation:** $\mathbb{E}[X] = \frac{1}{\lambda}$
2. **Variance:** $\mathrm{Var}(X) = \frac{1}{\lambda^2}$

This distribution is **memoryless**, i.e. $\mathbb{P}(X > s+t | X > s) = \mathbb{P}(X > t)$ when $X \sim \mathrm{Exp}(\lambda)$,$s, t > 0$

7.3 Gamma distribution

Idea: "Make exponential distribution more flexible"

Euler's gamma function is a mapping from $(0, \infty)$ to $\mathbb{R}$ defined by

$$\Gamma(a) = \int_0^\infty t^{a-1} \cdot e^{-t} \, dt$$

Proposition 7.1

1. *if $a > 0$, then $\Gamma(a+1) = a\Gamma(a)$*
2. *For $n \in \mathbb{N}$ follows $\Gamma(n) = (n-1)!$. In particular, $\Gamma(1) = \Gamma(2) = 1$ and $\Gamma(3) = 2$.*

Given $\alpha, \beta > 0$, a random variable $X \sim \Gamma(\alpha, \beta)$ if

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-x\beta} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Note: the exponential distribution with parameter λ is the $\Gamma(1, \lambda)$ distribution.

1. **Expectation:**
$$\mathbb{E}(x) = \frac{\Gamma(\alpha+1)\beta^\alpha}{\Gamma(\alpha)\beta^{\alpha+1}} = \frac{\alpha}{\beta}$$
2. **Variance:** $\mathrm{Var}(X) = \alpha/\beta^2$

7.4 Normal (or Gaussian) distribution

This section is devoted to the most important probability measure, the normal distribution. The idea is the "universal approximation for averages".
The normal distribution has a bell-shape density function and is used in the sciences to represent real-valued random variables that are assumed to be additively produced by many small effects.

Definition 7.2 *The probability measure generated by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is called normal distribution with expected value μ, and standard deviation σ. It is denoted by $\mathcal{N}(\mu, \sigma^2)$, that is, for all $a < b$

$$\mathcal{N}(\mu, \sigma^2)([a, b]) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

Definition 7.3 *The probability measure $\mathcal{N}(0, 1)$ is called standard normal distribution. It is given by*

$$\mathcal{N}(0, 1)([a, b]) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} \, dx$$

Proposition 7.4 *if $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ with $a \neq 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$*

Remark: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.

8 Random Vectors

Definition 8.1 *An 2-dimensional random vector is a function from a sample space Ω into $\mathbb{R}^2$.*

Definition 8.2 *If (X, Y) is a discrete random vector, the function*

$$f_{X,Y}(x, y) = \mathbb{P}((X, Y) = (x, y))$$

is called the ***joint probability mass function*** *of (X, Y)*

Key propery: $A \in \mathbb{R}^2$, then $\mathbb{P}((X, Y) \in A) = \sum_{(X,Y)\in A} f_{X,Y}(x, y)$

Definition 8.3 *A random vector (X, Y) is continuous if there exists a function $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$ such that*

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, dx\, dy$$

and its called ***the joint pdf*** *of (X, Y)*

Definition 8.4 *The joint cumulative distribution of the random vector (X, Y) is*

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$$

Note: in the continuous case, we have for continuous $f_{X,Y}$:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

Definition 8.5 *Let X be a continuous random variable and Y be a discrete random variable. Then, a function $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$ is called joint probability density function if*

$$\mathbb{P}((X, Y) \in A) = \int_{\mathbb{R}} \sum_{y:(x,y)\in A} f_{X,Y}(x, y)\, dx$$

Definition 8.6 ***marginal pmf*** *Let $f_{X,Y}(x, y)$ be a pmf for a random vector (X, Y), then*

$$f_X(x, y) = \sum_y f_{X,Y}(x, y)$$
$$f_Y(x, y) = \sum_x f_{X,Y}(X, Y)$$

are the marginal pmf respect to x and y respectively.

Definition 8.7 ***marginal pdf*** *Let $f_{X,Y}(x, y)$ be a pdf for a random vector (X, Y), then*

$$f_X(x, y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$
$$f_Y(x, y) = \int_{-\infty}^{\infty} f_{X,Y}(X, Y) dx$$

are the marginal pdf respect to x and y respectively.

8.1 Conditional distribution & independence

Definition 8.8 ***conditional pmf*** *Let (X,Y) be a discrete random vector with joint pmf $f_{X,Y}$ and marginals f_X and f_Y. The* ***conditional pmf*** *of X given Y is the function*

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

defined for all y such that $f_Y(y) \neq 0$

Definition 8.9 ***Conditional pdf*** *Let (X,Y) be continuous random variables with pdf $f_{X,Y}(x,y)$, then*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

is the conditional pdf of X given Y.

Remark: this definition holds also in the mixed case, i.e X continuous and Y discrete.

Definition 8.10 *Two random variables X, Y are independent if*

$$f_{X,Y} = f_X(x) \cdot f_Y(y)$$

both for discrete and continuous.

Proposition 8.11 ***Factorization criterion*** *Let two function $g, h : \mathbb{R} \to [0, \infty)$ such that*

$$f_{X,Y}(x,y) = g(x) \cdot h(y)$$

then X, Y are independent and

$$f_X(x) = \frac{g(x)}{\int_{-\infty}^{\infty} g(s)\, ds} \quad f_Y(y) = \frac{h(y)}{\int_{-\infty}^{\infty} g(t)\, dt}$$

Proposition 8.12 *Let X, Y be independent random variables and $A, B \in \mathbb{R}$. Then the events $\{X \in A\}$, $\{Y \in B\}$ are independent, i.e*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

Remark: the converse of this proposition holds too.

8.2 Expected value and variance

Theorem 8.13 *Let (X,Y) be random vector and and $g : \mathbb{R}^2 \to \mathbb{R}$, then*

$$\mathbb{E}[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y) \cdot f_{X,Y}(x,y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f_{X,Y}(x,y)\, dx\, dy & \text{if } X \text{ has pdf } f_X \end{cases}$$

Theorem 8.14 ***Linearity & Monotonility*** *Let (X,Y) be a random vector, then*

1. *let $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$*
2. *if $\mathbb{P}(X \geq Y) = 1$, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$*

Proposition 8.15 *If X and Y are independent, then for any g an d h we have*

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

in particular $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ *and* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

Definition 8.16 *If (X, Y) is a random vector, then*

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum_x x \cdot f_{X|Y}(x|y) & X \text{ discrete} \\ \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\, dx & X \text{ continuous} \end{cases}$$

is the conditional expectation of X given that $Y = y$

Note: $\mathrm{Var}(X|Y=y) = \mathbb{E}(X^2|Y=y) - (\mathbb{E}(X|Y=y))^2$

8.3 Transformation

Theorem 8.17 *Let X, Y be independent and $U = g_1(X)$, $V = g_2(Y)$ for some $g_1, g_2 : \mathbb{R} \to \mathbb{R}$. Then U, V are independent*

Theorem 8.18 ***pmf*** *Let (X_1, X_2) be a random vector and $g : \mathbb{R}^2 \to \mathbb{R}^2$ and set $(Y_1, Y_2) = g(X_1, X_2)$. Then,*

$$f_{Y_1,Y_2}(y_1, y_2) = \sum_{(x_1,x_2):g(x_1,x_2)=(y_1,y_2)} f_{X_1,X_2}(x_1, x_2)$$

Remark: Same works if $g : \mathbb{R}^2 \to \mathbb{R}$

Theorem 8.19 *if $X_1 \sim \mathrm{Poi}(\lambda_1)$ and $X_2 \sim \mathrm{Poi}(\lambda_2)$ are independent, then $X_1+X_2 \sim \mathrm{Poi}(\lambda_1+\lambda_2)$*

Theorem 8.20 ***pdf of*** $Y = g(X)$ *Let (X_1, X_2) be continuous random vector and g be differentiable with inverse $h(y) = g^{-1}(y)$. Then, for $(Y_1, Y_2) = g(X_1, X_2)$ we have*

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(h(y_1, y_2)) \cdot |J(y_1, y_2)|$$

where

$$J(y_1, y_2) = \det \begin{pmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{pmatrix}$$

8.4 Covariance and Correlation

Definition 8.21 *Let X, Y be random variables. Set $\mu_x = \mathbb{E}(X)$ and $\mu_y = \mathbb{E}(Y)$. Then,*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_x)(Y - \mu_y)]$$

*is the **covariance** of X and Y.*
Instead, set σ_X and σ_Y be the standard deviantion of X, Y respectively. Then,

$$\mathrm{Corr}(X, Y) = \rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

*is the **correlation** between X and Y*

Theorem 8.22 *1.* $\text{Cov}(X,Y) = \text{Cov}(Y,X)$

2. $\text{Cov}(X,X) = \text{Var}(X)$

3. $\rho_{X,X} = 1$

4. $\text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

5. X, Y *independent* $\rightarrow \text{Cov}(X,Y) = \rho_{X,Y} = 0$

Lemma 8.23 *For any random variable*

$$\mathbb{P}(X = 0) = 1 \Leftrightarrow \mathbb{E}(X^2) = 0$$

Theorem 8.24 ***Convariance properties***

1. $\text{Cov}(aX + bY, Z) = a\,\text{Cov}(X,Z) + b\,\text{Cov}(Y,Z)$

2. *if* X *or* Y *constant, then* $\text{Cov}(X,Y) = 0$

3. $\|\text{Cov}(X,Y)\| \leq \sigma_X\sigma_Y \rightarrow \text{Cov}(X,Y) \in [-1,1]$

4. *Assume* $\sigma_X, \sigma_Y > 0$. *Then,*

$$\text{Cov}(X,Y) = \sigma_X\sigma_Y \Leftrightarrow X = aY + b \text{ for some } a > 0, b \in \mathbb{R}$$
$$\text{Cov}(X,Y) = -\sigma_X\sigma_Y \Leftrightarrow X = aY + b \text{ for some } a < 0, b \in \mathbb{R}$$

Corollary 8.25 *Let* $\|\rho_{X,Y}\| \leq 1$ *and*

$$\rho_{X,Y} = 1 \Leftrightarrow Y = aX + b \quad , a > 0 \text{ "perfect correlation"}$$
$$\rho_{X,Y} = -1 \Leftrightarrow Y = aX + b \quad , a < 0 \text{ "perfect anti-correlation"}$$

Corollary 8.26 *1.* $\text{Cov}\left(\sum_{i\leq m} X_i, \sum_{j\leq n} Y_j\right) = \sum_{i\leq m}\sum_{j\leq n}\text{Cov}(X_i, Y_j)$

2. $\text{Var}\left(\sum_{i\leq n} X_i\right) = \sum_{i\leq n}\text{Var}(X_i) + 2\sum_{1\leq i<j\leq n}\text{Cov}(X_i, X_j)$

3. *If* $X_1, ..., X_n$ *are independent, then* $\text{Var}\left(\sum_{i\leq n} X_i\right) = \sum_{i\leq n}\text{Var}(X_i)$

9 Moment generating function

Definition 9.1 *The moment generating function of a random variable* X *is the function*

$$M_x(t) = \mathbb{E}(e^{tX}) := \begin{cases} \sum e^{tX} f_X(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tX} f_X(x) & X \text{ continuous} \end{cases}$$

provided that the sum/integral converges for all t *in an interval of the form* $(-h,h), h > 0$.

Proposition 9.2 ***Linearity*** *Let* X *a random variable, and* $a, b \in \mathbb{R}$, *then*

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

Theorem 9.3 *If X, Y are such that $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X = F_Y$ (that is, X and Y have the same distribution).*

Proposition 9.4 *If X and Y are independent, then*

$$M_{X+Y}(t) = M_X(t)M_Y(t) \quad \text{for all } t \geq 0.$$

10 The bivariate normal distribution

Definition 10.1 *Let (X, Y) be a random vector. We say that (X, Y) is **bivariate normal** with parameters $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$ and $\rho \in (-1, -1)$ if it has joint pdf*

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y}\right)\right\}$$

we write: $(X, Y) \sim \mathcal{N}((\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$

Lemma 10.2 ***Existence*** *Let $Z_1 \sim \mathcal{N}(0, 1)$, $Z_2 \sim \mathcal{N}(0, 1)$ be independent. Set*

$$U := \sigma_1 Z_1 + \mu_1$$
$$V := \rho\sigma_2 Z_1 + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2$$

then

$$(U, V) \sim \mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

Proposition 10.3 *If $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then*

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), \quad \rho_{X,Y} = \rho.$$

Corollary 10.4 *If $(X, Y) \sim \mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, then*

$$aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y).$$

11 Higher dimensions

Definition 11.1 ***Random vector*** *An n-dimensional random vector is a function from a sample space Ω into $\mathbb{R}^n$.*

Definition 11.2 ***(Joint pmf)*** *If $(X_1, ..., X_n)$ is a discrete random vector, the function*

$$f_{X_1,...,X_n}(x_1, ..., x_n) := \mathbb{P}((X_1, ..., X_n) = (x_1, ..., x_n))$$

is called the joint probability mass function (pmf) of $(X_1, ..., X_n)$ (sometimes we omit the word "joint" and simply say that $f_{X_1,...,X_n}$ is the pmf of the random vector).

Remark: $\mathbb{P}((X_1, ..., X_n) \in A) = \sum_{(x_1,...,x_n)\in A} f_{X_1,...,X_n}(x_1, ..., x_n)$

Definition 11.3 ***Joint pdf*** *A random vector* $(X_1, ..., X_n)$ *is continuous if there exists a function* $f_{X_1,...,X_n}$ *such that*

$$\mathbb{P}((X_1, ..., X_n) \in A) = \int \cdots \int_A f_{X_1,...,X_n}(x_1, ..., x_n)\, dx_1 \cdots dx_n$$

$f_{X_1,...,X_n}$ *is called the joint probability density function (pdf) of* $(X_1, ..., X_n)$*.*

Definition 11.4 ***Joint cdf*** *The joint cumulative distribution function (cdf) of the random vector* $(X_1, ..., X_n)$ *is*

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \mathbb{P}(X_1 \leq x_1, \cdots, X_n \leq x_n)$$

In the continuous case, we have:

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \frac{\partial^n F_{X_1,...,X_n}}{\partial x_1 \cdots \partial x_n}(x_1, ..., x_n)$$

Definition 11.5 ***Expectation*** *If* $(X_1, ..., X_n)$ *is a random vector and* $g : \mathbb{R}^n \to \mathbb{R}$*, then*

$$\mathbb{E}(g(X_1, ..., X_n)) = \begin{cases} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, ..., x_n) f_{X_1,...,X_n}(x_1, ..., x_n)\, dx_1 \cdots dx_n & \text{in the continuous case} \\ \sum_{x_1} \cdots \sum_{x_n} g(x_1, ..., x_n) f_{X_1,...,X_n}(x_1, ..., x_n) & \text{in the discrete case} \end{cases}$$

Definition 11.6 ***Conditional pmf/pdf*** *Le* $(X_1, ..., X_n)$ *be a continuous/discrete random vector. The conditional pmf/pdf of* $(X_1, ..., X_m)$ *given* $(X_{m+1}, ..., X_n)$ *is*

$$f_{X_1,...,X_m|(X_{m+1},...,X_n}(x_1, ..., x_m|x_{m+1}, ..., x_n) = \frac{f_{X_1,...,X_n}(x_1, ..., x_n)}{f_{X_{m+1},...,X_n}(x_{m+1}, ..., x_n)}$$

defined for all $x_1, ..., x_m$ *and for all* $x_{m+1}, ..., x_n$ *such that* $f_{X_{m+1},...,X_n}(x_{m+1}, ..., x_n) > 0$*.*

Definition 11.7 ***Independence*** *The random variables* $X_1, ..., X_n$ *are independent if*

$$f_{X_1,...,X_n}(x_1, ..., x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

(both for discrete and continuous).

Definition 11.8 ***Joint mgf*** *The joint mgf of a random vector* $(X_1, ..., X_n)$ *is the function* $M_{X_1,...,X_n}(t_1, ..., t_n) = \mathbb{E}(e^{t_1 X_1 + \cdots + t_n X_n})$*. That is*

$$M_{X_1,...,X_n}(t_1, ..., t_n) = \begin{cases} \sum e^{t_1 X_1 + \cdots + t_n X_n} \cdot f_{X_1,...,X_n}(x_1, ..., x_n) & (X_1, ..., X_n) \text{ discrete} \\ \int_{-\infty}^{\infty} e^{t_1 X_1 + \cdots + t_n X_n} \cdot f_{X_1,...,X_n}(x_1, ..., x_n)\, dx_1 \cdots dx_n & (X_1, ..., X_n) \text{ continuous} \end{cases}$$

provided that the sum/integral converges in an interval of the form $(-h, h)^n$ $h > 0$

12 Statistic

Definition 12.1 ***Random Sample*** *A random sample of size* n *is a sequence* $X_1, ..., X_n$ *of independent random variables all with the same pdf/pmf, say say* $f(x)$*. We thus have*

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{1 \leq i \leq n} f_{X_i}(x_i)$$

we say that f *is the* ***population pdf/pmf***

Remark:

1. There is an infinite population of some entities
2. Each entity has some attribute
3. f describes attribute distribution over the population
4. We select n individuals and record their attributes to obtain $X_1, ..., X_n$

Definition 12.2 ***Parameter*** *A parameter is a constant that defines the population pmf/pdf* $f(x)$

Definition 12.3 ***Statistic*** *A statistic is a function* $T : \mathbb{R}^n \to \mathbb{R}$*of a random sample.*

$$Y = T(X_1, ..., X_n)$$

Definition 12.4 *A statistic* Y *is a unbiased estimator for the parameter* θ $\mathbb{E}(Y) = \theta$

Definition 12.5

$$\begin{aligned} \textit{sample mean} &: \quad \overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \\ \textit{sample variance} &: \quad S_n^2 = \frac{1}{n-1} \sum_{i \leq n} (X_i - \overline{X})^2 \end{aligned}$$

Lemma 12.6

$$S_n^2 = \frac{1}{n-1} \sum_{i \leq n} X_i^2 - \frac{n}{n-1} \overline{X}_n^2$$

Theorem 12.7 ***Unbiasedness of sample mean variance*** *Let* $X_1, ..., X_n$ *be independent and identically distributed with mean* μ *and variance* σ^2*. Then,*

1. $\mathbb{E}(\overline{X}_n) = \mu$
2. $\mathbb{E}(S_n^2) = \sigma^2$

12.1 Convergence concepts

The idea is how large should n be such that $\overline{X}_n$ approximates μ well?

Definition 12.8 *A sequence of $X_1, X_2, ...$ of random variables converges in probability to a constant $c \in \mathbb{R}$ if $\forall \epsilon > 0$:*

$$\mathbb{P}(|X_n - c| > \epsilon) \to 0$$

we write $X_n \xrightarrow[n\to\infty]{\mathbb{P}} c$

Definition 12.9 *Let $X_1, ..., X_n$ be a random sample of pmf/pdf with parameter θ. We say that Y_n is* ***consistent estimator*** *of θ if*

$$Y_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

Theorem 12.10 ***Weak Law of Large Numbers*** *Let $X_1, X_2, ...$ independent and identically distributed with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$ then*

$$\overline{X}_n \underset{\mathbb{P}}{\to} \mu \qquad \lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = 0$$

Definition 12.11 ***converges in distribution*** *A sequence of random variables $X_1, X_2, ...$ converges in distribution to a random variable X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for every $x \in \mathbb{R}$ at which $F_X(x)$ is continuous. We denote this by

$$X_n \xrightarrow[d]{n\to\infty} X$$

Lemma 12.12 *If X is continuous and $X_n \xrightarrow[d]{n\to\infty} X$, then*

$$\mathbb{P}(X_n = x) \xrightarrow{n\to\infty} 0$$

for all $x \in \mathbb{R}$

Proposition 12.13 *If X is continuous and $X_n \xrightarrow[d]{n\to\infty} X$, then for every interval $I \subset \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}(X_n \in I) = \mathbb{P}(X \in I)$$

Theorem 12.14 ***Central Limit Theorem*** *Let $X_1, X_2, ...$ be independent and identically distributed with mean μ and variance σ^2 (both finite). Then,*

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \xrightarrow[d]{n\to\infty} Z, \qquad \text{where } Z \sim \mathcal{N}(0,1)$$

Remarks:

$$\sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} = \frac{\sum_{i=1}^{n} X_i - \mu n}{\sigma\sqrt{n}}$$

Theorem 12.15 *Assume that* $X_1, X_2, ...$ *and* X *are such that*

$$M_{X_n}(t) \xrightarrow{n\to\infty} M_X(t)$$

or all t *in a neighborhood of* 0. *Then,* $X_n \xrightarrow[d]{n\to\infty} X$

Theorem 12.16

$$\mathbb{E}(X^n) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

Theorem 12.17 ***Normal approximation to binomial*** *When n is large and p is not too close to 0 or 1, we have the approximation*

$$X \sim \text{Bin}(n,p) \approx Y \sim \mathcal{N}(np, np(1-p))$$

where

$$\mathbb{P}(X \le b) \approx \int_{-\infty}^{b+\frac{1}{2}} f_Y(y)\,dy = F_Y\left(b+\frac{1}{2}\right), \quad \mathbb{P}(X \ge a) \approx \int_{a-\frac{1}{2}}^{\infty} f_Y(y)\,dy = 1 - F_Y\left(a-\frac{1}{2}\right)$$

this approximation holds if $n \ge 15$, $np \ge 5$ *and* $n(1-p) \ge 5$.

Theorem 12.18 ***Chebyshev Inequality*** *Let* X *an random variable,*

$$\mathbb{P}(|X - \mathbb{E}(X)| > x) \le \frac{\text{Var}(X)}{x^2}, \quad x > 0$$

13 Random Walk

Definition 13.1 $X_1, X_2, ...$ *independent random variables with values in* $\{-1, 1\}$ *set* $p := \mathbb{P}(X_1 = 1)$, $q := \mathbb{P}(X_1 = -1)$ *and set* $S_0 \ge 0$. *Then, the sequence*

$$S_n := S_0 + X_1 + X_2 + \cdots + X_n$$

is called ***simple random walk***

Theorem 13.2 ***pmf of*** S_n *Suppose that* $n + k$ *is even. Then,*

$$\mathbb{P}(S_n - S_0) = \binom{n}{\frac{n+k}{2}} p^{\frac{n+k}{2}} q^{\frac{n-k}{2}}$$

Definition 13.3 ***Passage times*** *Let* $\{S_n\}_{n\ge0}$ *be a simple random walk with* $S_0 = i$. *Then,*

$$T_{i,k} := \min\{n \ge 1 : S_n = k\}$$

is the passage time from i *to* k

Theorem 13.4 ***Finiteness criterion***

$$\mathbb{P}(T < \infty) = \begin{cases} 1 & p \ge q \\ \frac{p}{q} & p < q \end{cases}$$

Theorem 13.5 ***Finite expected passage time*** *If* $p > q$*, then* $\mathbb{E}[T] < \infty$*.*

Theorem 13.6 ***Markov inequality*** *Let* $a > 0$ *and* Y *be any non-negative random variable. Then,*

$$\mathbb{P}(Y \geq a) \leq \frac{1}{a}\mathbb{E}(Y)$$

Theorem 13.7 ***pmf of*** $T_{0,b}$ *Let* $(S_n)_{n\geq 0}$ *be simple random walk with* $S_0 = 0$*. Then, for* $b > 0$*,*

$$\mathbb{P}(T_{0,b} = n) = \frac{b}{n}\mathbb{P}(S_n = b)$$

References

[1] Christian Hirsch. Probability theory. 2021.

[2] Werner Linde. *Probability Theory*. De Gruyter, 2016.